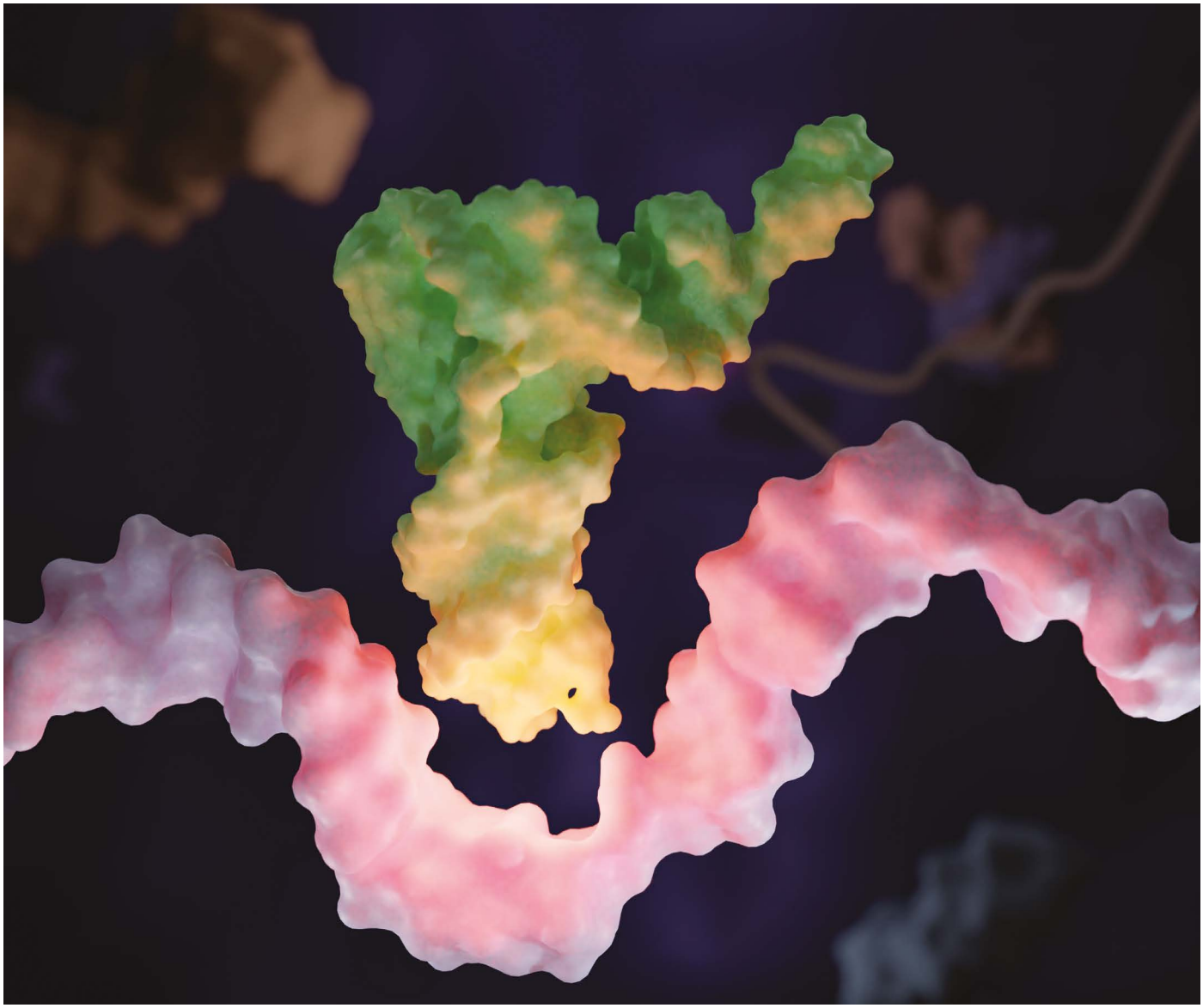


GETTY

RNA structure poses unique challenges for computational models.

WHY IS RNA STRUCTURE SO HARD TO PREDICT?

AlphaFold's highly accurate structural models transformed protein biology, but RNA prediction lags behind. **By Diana Kwon**

At a virtual conference in November 2020, the winner of a biennial protein-structure-prediction challenge was announced: AlphaFold. Created by Google DeepMind, this computational tool had blown its competitors out of the water by solving dozens of protein structures with atomic-level accuracy, accomplishing a feat that researchers had been attempting for decades.

The challenge, known as the Critical Assessment of Protein Structure Prediction (CASP), was launched in 1994 to advance computational tools for modelling 3D protein configurations from their amino-acid sequences. Teams of scientists pitted their computational models against each other, trying to generate the most accurate predictions for previously unknown protein structures that, before the event, are solved experimentally using methods such as X-ray crystallography and cryo-electron microscopy.

AlphaFold's 2020 predictions rivalled those solved with these tried-and-tested techniques, and it has since become a favourite of the structural-biology community. Its repository — the AlphaFold Protein Structure Database — contains some 200 million structures, and, in 2024, AlphaFold's developers shared half of the Nobel Prize in Chemistry for their work.

But that's proteins. In 2022 the organizers of CASP turned their attention towards a different, yet still challenging, class of biomolecules: RNA.

As with proteins, determining RNA structure typically requires costly and time-consuming experimental methods. Computational tools can help, but RNA is a tougher nut to crack. One simple reason, according to Yu Li, a computer scientist at the Chinese University of Hong Kong, is historical. For a long time, most scientists didn't think RNA biology was interesting enough to study. But RNA also poses unique molecular challenges, and relatively few data are available to train computational models of the type that have performed so well with proteins.

Researchers have been getting creative, however, and there is a growing toolkit of computational tools emerging to aid the prediction of RNA structure. Many of these incorporate the latest developments in artificial intelligence (AI), including the large language models (LLMs) that underlie popular chatbots, such as ChatGPT.

"RNA folding is a very tough problem," concedes Shi-Jie Chen, a computational biophysicist at the University of Missouri in Columbia. But AI, he adds, is getting "better and better".

Elusive targets

For a long time, RNA was seen simply as an intermediary between two more interesting classes of molecule: DNA, the 'blueprint of life', and proteins, the 'building blocks' of the cell. Only a small fraction of the human genome encodes proteins, yet much of the non-coding genome is transcribed into RNA. Over the past few decades, scientists have discovered that these non-coding RNAs mediate essential functions in healthy cells — and contribute to many diseases.

How these RNAs work remains, in many cases, a mystery. Researchers hope that, by determining their shape, they will be able to understand better the role that these molecules have in making our cells tick — a question of form dictating function. "In biology, we assume that the sequence is very likely to determine the structure, and that the structure is very likely to determine the function," says Li.

But computational tools for predicting RNA structure lag behind their protein equivalents. Even AlphaFold3, the latest version of DeepMind's structure-prediction tool — falls short when it comes to RNA.

"If you look at the recent CASP competitions, we are at the point where, on the protein structure side of things, fully automated teams are as good as human teams," says Lydia Freddolino, a systems biologist at the University of Michigan in Ann Arbor and a scientific advisory board member for CircNova, a company that uses deep-learning tools to design circular RNA-based therapeutics. "For RNA, we are nowhere near that — all the top groups make heavy use of human intervention."

"Segments of RNA interact in all kinds of weird and wonderful ways."

RNA-structure prediction featured in CASP competitions in 2022 and 2024, and Freddolino participated in both. The team that ranked first for predicting RNA structures at the latest event, CASP16, used a hybrid approach: combining AI with a defined, physics-based algorithm. According to Chen, who led the winning group, they first used AlphaFold3 to generate ensembles of possible RNA structures, and then applied a physics-based model that probes the 'energy landscape' of possible structures to pinpoint the conformations that are most likely to form. (Chen's team has licensed their software to several biotechnology firms.)

Researchers developing AI-only tools for predicting RNA structure face numerous obstacles. One is that RNA molecules have features that make their structures inherently hard to predict. RNA molecules have more flexible backbones than do proteins, and their structures are more dynamic, meaning that they can undergo substantial conformational changes while carrying out their biological tasks.

On top of that, RNA molecules lack the different chemistries that can be found in proteins, such as acidic and basic residues, that allow for stable connections to form. Instead, segments of RNA interact in all kinds of "weird and wonderful ways", Freddolino says, such as through different base pairings and the involvement of metal ions. As a result, the subtle variations between the best and worst models are trickier to spot than with proteins.

The chemical alphabet of RNA is also harder to interpret: the four chemical bases that make up RNA are less distinct than the 20 amino acids found in proteins. That means that each RNA base contains less information than an amino acid. One reason tools such as AlphaFold have been so successful, Freddolino notes, is the ability to use large sequence databases to pinpoint patterns of interactions between different amino acids — and this is much more difficult to do with RNA.

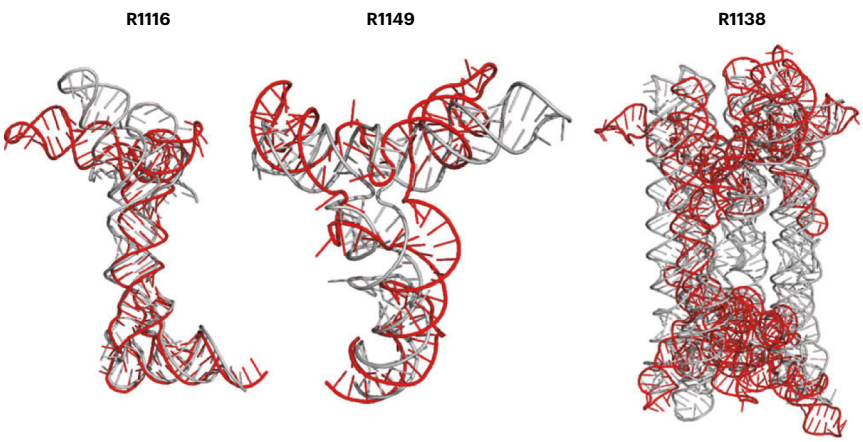
And then there's the paucity of known RNA structures. The Protein Data Bank, a repository of 3D macromolecular structures, contains nearly 200,000 protein structures and fewer than 2,000 RNAs. This dearth of data means that there is less information to feed the algorithms that underlie AI-based structure prediction.

"We're doing as well as we can with the limited data that we have," says Jim Collins, a biomedical engineer at the Massachusetts Institute of Technology in Cambridge. "The field would advance considerably with the collection and curation of many more structures."

Bringing in AI

Researchers have been working to address these challenges, and, in recent years, several AI-based RNA-structure-prediction tools have emerged. Before 2020, most of the methods for predicting RNA structure were based on algorithms defined by specific physical or mathematical models, according to Jianyi Yang, a computational biologist at Shandong University in Qingdao, China. But the success of AlphaFold has inspired people in the RNA field to apply AI to this problem, too, he says.

Yang and his colleagues designed a fully automated (and freely available) AI tool, trRosettaRNA, which combines deep learning with elements of Rosetta, a computational tool used for determining molecular structures that was created by David Baker at the University of Washington in Seattle, who shared



W. WANG *ET AL./NATURE COMMUN.*

Natural (R1116 and R1149) and synthetic (R1138) RNA structures, used in the structure-prediction task CASP15, measured experimentally (grey) and predicted using an AI tool (red).

the 2024 chemistry Nobel with the creators of AlphaFold.

Just as for proteins, the structure of RNA occurs on multiple levels: nucleotide sequence (primary); intermediary structures that form when base pairs find their complements (secondary); and the final, 3D structure (tertiary). RNAs can also form complexes with each other and other molecules (quaternary). First, trRosettaRNA generates predictions of primary and secondary structures, then, with the help of a classical physics-based model, it reconstructs tertiary structures. Secondary structures — such as 'hairpins' that form when short segments of sequence pair up with one another — are much more important for RNA than they are for proteins, Yang says, and using these in-between structures is one of the keys to this model's success.

Yang's team pitted trRosettaRNA against other automated tools and found, on the basis of an assessment with two independent data sets of dozens of RNAs, that it surpassed those tools in accuracy[1]. In 2024, the software placed fourth at CASP16.

Other teams are applying LLMs to solve RNA structures. Li and his colleagues, including Collins, developed one such tool, called RhoFold, which infers information about RNA structure directly from its sequence[2]. One of the core assumptions of RhoFold, Li explains, is that nucleotides that are near to one another in 3D space have co-evolved to form distinct patterns in RNA sequences. To help their LLM to find these patterns, it was trained on some 23 million RNA sequences with 'masked' regions so that it can learn to identify patterns in unmasked segments — which are, in turn, used to predict what the hidden sequences are, he explains.

In November 2024, Li and his team reported[2] that RhoFold could accurately predict RNA structures, outperforming some of the top scorers of the 2022 CASP competition on many RNA targets, including teams that had experts in the pipeline. RhoFold, which is available for free online, can be used in a variety of applications, Li says, such as understanding the interactions between RNA and other molecules — and generating candidate structures for drug design.

Li and his colleagues demonstrated the latter application in a second paper published last year[3]. They integrated RhoFold into a pipeline for another tool, RhoDesign, which uses generative AI to design new RNA sequences. Rather than trying to understand function through structure, RhoDesign enables scientists to design an RNA sequence that leads to a desired structure — and these RNAs can then be tested in the lab to determine whether they function in the expected way. By integrating RhoFold, they were able to rapidly evaluate the potential of these structures by comparing them with existing RNAs.

Using this pipeline, Li's team produced RNA aptamers (molecules that can recognize and bind to targets with a high degree of affinity and specificity) that fluoresce in the presence of specific small molecules. They say that this provides proof of principle that the approach can be used to develop more interesting aptamers, including therapeutics or diagnostics. The code for RhoDesign is available online.

Still other teams are developing tools to generate artificial RNA sequences from scratch with specific, designed functions. In January 2024, Hirohide Saito, a bioengineer at Kyoto University in Japan, and his colleagues showed that they could use a generative AI tool called RfamGen to create functional ribozymes — enzymes built of RNA rather than protein — that could catalyse reactions more efficiently than do naturally occurring ones[4]. Among the ribozymes they created are 'riboswitch ribozymes', which cleave their own sequences in the presence of specific ligands.

> **"We're doing as well as we can with the limited data that we have."**

A different approach

And there are still other applications of AI to RNA biology. Jamie Cate, a structural biologist at the University of California, Berkeley, and his colleagues developed a generative AI model that predicts mutations that improve RNA function by linking RNA sequences with existing data on how they behave. In one study[5], they showed that their model — when trained on RNA sequences and the optimal growth temperatures of the organisms they originated from — could be used to pinpoint mutations that increased the ability of the ribosomes to withstand high temperatures. This work illustrates the feasibility of using AI to aid in the process of genetically engineering new functions into ribosomes, Cate and his coauthors say — a key area of interest for many scientists.

Biotechnology firm Atomic AI in San Francisco, California, meanwhile, has developed a platform for using computational RNA structure prediction to help design potential therapeutics. This platform is based, in part, on a deep-learning model called ATOM-1, which incorporates data from chemical probing, an experimental approach in which segments of an RNA are exposed to molecules that react to specific structural conformations, such as paired or unpaired bases[6]. These molecules can induce chemical changes in the RNA — and, because these modifications are structure dependent, this provides 'indirect' information about the RNA's structure, explains Stephan Eismann, the company's founding scientist and head of machine learning. These AI-based models help generate new targets, which the company's scientists use as part of the decision-making process, Eismann says.

But what's really needed are data — and lots of them — to feed the AI tools. As well as conducting further studies with tried-and-tested structural techniques such as cryo-electron microscopy, researchers have devised methods to collect these data in other ways. For example, Rhiju Das, a biochemist at Stanford University in California, and his colleagues created Eterna, an online game that allows anyone to help design RNAs by solving online puzzles. These crowdsourced RNA designs are then tested in the lab and returned to players to improve their predictions. Using this platform, Das's team has created a data set of chemical-mapping measurements for two million RNA sequences, dubbed Ribonanza. Through a subsequent competition on the machine-learning-challenge website Kaggle, the team developed RibonanzaNet, a deep-learning model that could predict secondary RNA structures[7].

At the end of February, Das and his colleagues, in collaboration with the organizers of CASP and another popular RNA-structure-prediction competition, RNA-Puzzles, launched a new Kaggle competition that challenges researchers to create computational methods of RNA 3D structure prediction capable of outperforming experts. This contest requires participants to create fully automated models, and Das is hopeful that someone will be able to achieve this goal.

Still, experts say that much more work is needed before RNA structures can be predicted with the same level of sophistication as is possible for proteins. And whenever RNA researchers do reach what some in the field call the 'AlphaFold moment' — the point at which RNA-structure prediction reaches the accuracy of which AlphaFold is capable — experts stress that predictions are just that. They still need to be tested in the lab. But, if AlphaFold is "not a panacea", Freddolino says, it's provided a powerful tool for structural biologists — and researchers hope to get there with RNA as well.

"Functional RNAs are hugely important in all domains of life," Freddolino says. "And we are, comparatively, in our infancy in terms of our ability to predict their structure."

1. Wang, W. *et al*. *Nature Commun.* **14**, 7266 (2023).
2. Shen, T. *et al*. *Nature Methods* **21**, 2287–2298 (2024).
3. Wong, F. *et al*. *Nature Comput. Sci.* **4**, 829–839 (2024).
4. Sumi, S., Hamada, M. & Saito, H. *Nature Methods* **21**, 435–443 (2024).
5. Shulgina, Y. *et al*. *Nature Commun.* **15**, 10627 (2024).
6. Boyd, N. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2023.12.13.571579 (2023).
7. He, S. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2024.02.24.581671 (2024).